Data Mining CSE 572- Assignment 4 / Mini Project 2 Submission Report

Ankit Nadig (1211213650), anadig@asu.edu
Vraj Delhivala (1211213637), vdelhiva@asu.edu

Problem Statement- Study the application of the k-means clustering algorithm and active learning for classification problems.

Problem 1

1. First, we import the seeds.txt file into a Matrix
2. For each value of k ie. 3,5 and 7: we randomly initialise k number of centroids 10 times.
3. In each of the 10 random initializations, we iterate till convergence. In each iteration, we assign all points in the dataset to the centroid closest to the point and then we update the centroids by updating the centroid value to be the mean of all points in the cluster, the centroid belongs to. We then calculate the sum of squared error and calculate the change in the SSE.
4. We stop at 100 iterations or when the sum of squared error changes by less than 0.001

K	Average Sum of Squared Error
3	587.903964
5	406.286258
7	294.802082

Average SSE value (averaged over 10 initializations) for Problem1. This value can change due to the Random data sampling

Problem 2

Here, we have three sets of data(training,testing and unlabeled) in each dataset of Mind Reading and MMI. We load one set of data from Mind Reading / MMI and do the following operations -

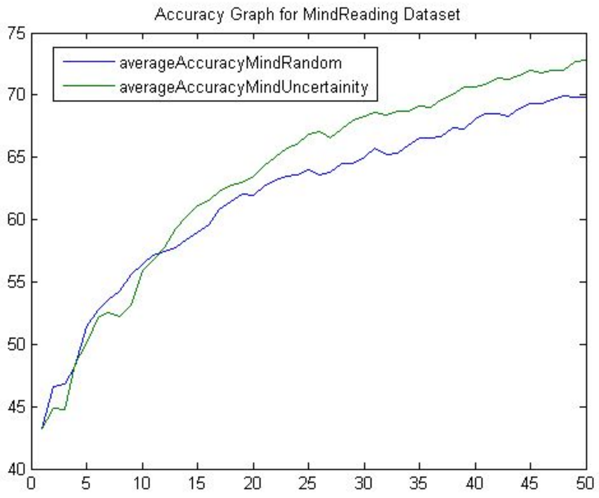
a. We pass this data to our Random Based Active Learning function and do the following over 50 iterations -
    i. We train the classifier by using the training data and the given function, ‘train_LR_Classifier’ to get the output of trained_weights.
    ii. Using these weights,we find the probabilities for the test data using the ‘test_LR_Classifier’ function and determine the accuracy.
    iii. We use the ‘test_LR_Classifier’ now on the unlabeled data.

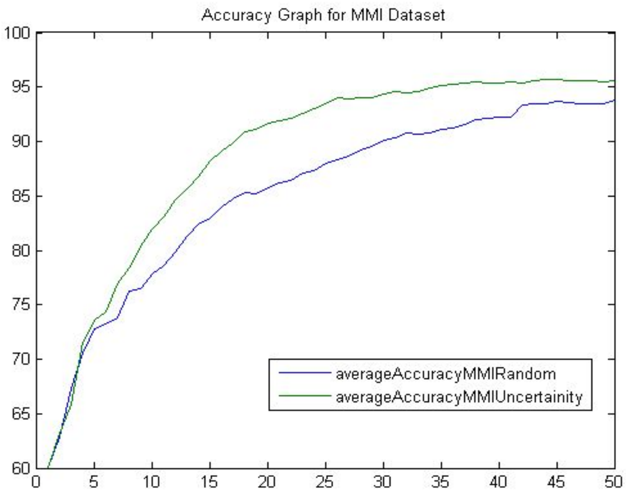
iv. Next, we pick up Random Samples from the Unlabeled dataset and add them to the training set while removing those samples from the unlabeled set.
v. Doing this, we get the accuracy over each iteration.

b. We pass this data to our Entropy Based Active Learning function and do the following over 50 iterations -
   i. We train the classifier by using the training data and the given function, ‘train_LR_Classifier’ to get the output of trained_weights.
   ii. Using these weights,we find the probabilities for the test data using the ‘test_LR_Classifier’ function and determine the accuracy.
   iii. We use the ‘test_LR_Classifier’ now on the unlabeled data and get the probability vector which we use to calculate uncertainty measure for each sample(row) and sort the data sample based on entropy(descending)
   iv. We pick top 10 samples with highest entropy from the Unlabeled dataset and add them to the training set while removing those samples from the unlabeled set.
   v. Doing this, we get the accuracy over each iteration.

Following these two steps for all three sets of data in both the Data sets we get values for accuracy that we average out to get Average Accuracy values for both data sets and both Active learning methods over 50 iterations.

Following are the Accuracy Graphs (Higher resolution images in the Submission Folder)


Accuracy Graph for MindReading Dataset
averageAccuracyMindRandom
averageAccuracyMindUncertainity


Accuracy Graph for MMI Dataset
averageAccuracyMMIRandom
averageAccuracyMMIUncertainity